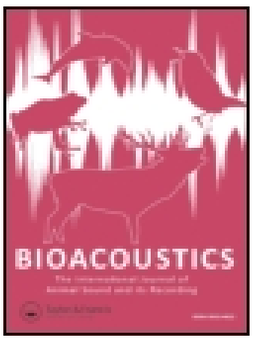
Who shall I say is calling? Validation of a caller recognition procedure in Bornean flanged male orangutan (Pongo pygmaeus wurmbii) long calls

Spillmann, Brigitte ; van Schaik, Carel P ; Setia, Tatang M ; Sadjadi, Seyed Omid

Abstract: Acoustic individual discrimination has been demonstrated for a wide range of animal taxa. However, there has been far less scientific effort to demonstrate the effectiveness of automatic individual identification, which could greatly facilitate research, especially when data are collected via an acoustic localization system (ALS). In this study, we examine the accuracy of acoustic caller recognition in long calls (LCs) emitted by Bornean male orangutans (Pongo pygmaeus wurmbii) derived from two data-sets: the first consists of high-quality recordings taken during individual focal follows (N = 224 LCs by 14 males) and the second consists of LC recordings with variable microphone-caller distances stemming from ALS (N = 123 LCs by 10 males). The LC is a long-distance vocalization. We therefore expect that even the low-quality test-set should yield caller recognition results significantly better than by chance. Automatic individual identification was accomplished using software originally developed for human speaker recognition (i.e. the MSR identity toolbox). We obtained a 93.3% correct identification rate with high-quality recordings, and 72.23% with recordings stemming from the ALS with variable microphone-caller distances (20–420 m). These results show that automatic individual identification is possible even though the accuracy declines compared with the results of high-quality recordings due to severe signal degradations (e.g. sound attenuation, environmental noise contamination, and echo interference) with increasing distance. We therefore suggest that acoustic individual identification with speaker recognition software can be a valuable tool to apply to data obtained through an ALS, thereby facilitating field research on vocal communication.

DOI: https://doi.org/10.1080/09524622.2016.1216802



Bioacoustics
The International Journal of Animal Sound and its Recording

ISSN: 0952-4622 (Print) 2165-0586 (Online) Journal homepage: http://www.tandfonline.com/loi/tbio20

Who shall I say is calling? Validation of a caller recognition procedure in Bornean flanged male orangutan (Pongo pygmaeus wurmbii) long calls

Brigitte Spillmann, Carel P. van Schaik, Tatang M. Setia & Seyed Omid Sadjadi

To cite this article: Brigitte Spillmann, Carel P. van Schaik, Tatang M. Setia & Seyed Omid Sadjadi (2016): Who shall I say is calling? Validation of a caller recognition procedure in Bornean flanged male orangutan (Pongo pygmaeus wurmbii) long calls, Bioacoustics, DOI: 10.1080/09524622.2016.1216802

To link to this article: http://dx.doi.org/10.1080/09524622.2016.1216802

BIOACOUSTICS, 2016
http://dx.doi.org/10.1080/09524622.2016.1216802

Taylor & Francis
Taylor & Francis Group

Who shall I say is calling? Validation of a caller recognition procedure in Bornean flanged male orangutan (*Pongo pygmaeus wurmbii*) long calls

Brigitte Spillmann[a], Carel P. van Schaik[a], Tatang M. Setia[b] and Seyed Omid Sadjadi[c]

[a]Anthropological Institute and Museum, University of Zurich, Zurich, Switzerland; [b]Fakultas Biologi, Universitas Nasional, Jakarta, Indonesia; [c]IBM Research, Yorktown Heights, NY, USA

ABSTRACT

Acoustic individual discrimination has been demonstrated for a wide range of animal taxa. However, there has been far less scientific effort to demonstrate the effectiveness of automatic individual identification, which could greatly facilitate research, especially when data are collected via an acoustic localization system (ALS). In this study, we examine the accuracy of acoustic caller recognition in long calls (LCs) emitted by Bornean male orangutans (*Pongo pygmaeus wurmbii*) derived from two data-sets: the first consists of high-quality recordings taken during individual focal follows ($N = 224$ LCs by 14 males) and the second consists of LC recordings with variable microphone-caller distances stemming from ALS ($N = 123$ LCs by 10 males). The LC is a long-distance vocalization. We therefore expect that even the low-quality test-set should yield caller recognition results significantly better than by chance. Automatic individual identification was accomplished using software originally developed for human speaker recognition (i.e. the MSR identity toolbox). We obtained a 93.3% correct identification rate with high-quality recordings, and 72.23% with recordings stemming from the ALS with variable microphone-caller distances (20–420 m). These results show that automatic individual identification is possible even though the accuracy declines compared with the results of high-quality recordings due to severe signal degradations (e.g. sound attenuation, environmental noise contamination, and echo interference) with increasing distance. We therefore suggest that acoustic individual identification with speaker recognition software can be a valuable tool to apply to data obtained through an ALS, thereby facilitating field research on vocal communication.

ARTICLE HISTORY
Received 3 May 2016
Accepted 9 July 2016

KEYWORDS
Caller recognition; mel-frequency cepstral coefficients; Gaussian mixture model; acoustic localization system (ALS); long call; orangutan

Introduction

Acoustic individual recognition can be a very useful tool in studies using passive acoustic monitoring (PAM) for applications such as censuses or acoustic localization systems (ALSs) for tracking individuals. The presence of individually distinctive acoustic characteristics in

CONTACT Brigitte Spillmann brigitte.spillmann@uzh.ch

calls or songs has been demonstrated for a wide range of animal taxa, from anurans (Bee et al. 2001) and birds (Ehnes and Foote 2015) to mammals (Townsend et al. 2014), including non-human primates (Wich et al. 2003). These differences are used for individual recognition in a wide range of animal taxa and modalities, as indicated by playback experiments (Cheney and Seyfarth 1982; Rendall et al. 1996), stressing the role of individual recognition in moulding social behaviour (Tibbetts and Dale 2007). It might therefore be beneficial for studies using PAM to extract caller identities in order to address biologically meaningful research questions about the social behaviour of the study species, such as territoriality, aggressive competition, mate attraction, etc. Nevertheless, although most of the studies addressing vocal individuality show that it is possible to discriminate among individuals, it has turned out to be quite challenging to achieve reliable identification of individuals.

Two classes of feature extraction approaches are in use for individual recognition: statistical and non-statistical (Cheng et al. 2010). The former, commonly used for animal calls, relies on measurements of the acoustic call structure from waveform or spectrographic representations of recorded calls (Galeotti and Sacchi 2001; Peake and McGregor 2001; Gilbert et al. 2002; Kirschel et al. 2009; Xia et al. 2012). The latter class, which is commonly adopted in human speaker recognition systems, includes linear prediction-based cepstral coefficients and mel-frequency cepstral coefficients (MFCCs). In the MFCC feature extraction procedure (see Figure 1), on which we focus here, acoustic waveforms are transformed into compact feature vectors on a frame-by-frame basis taking into account an approximate model of human's auditory perception. This auditory perception model is reflected in the mel-frequency scale (Davis and Mermelstein 1980). An example of the auditory-inspired mel-scale filterbank with 27 channels is provided in Figure 2. This cepstral representation captures the vocal tract resonances, and is based on the source-filter model of human speech production, which is also used to describe the vocal production system in many animal species (Fitch 2006; Taylor and Reby 2010). Compared with MFCC extraction, manual acoustic feature extraction is time-consuming and influenced by the researcher's intuition-based decisions on which parameters to extract. In contrast, MFCC extraction is fully automated, repeatable, and standardized (Mielke and Zuberbühler 2013).

Figure 1. Schematic block diagram of the MFCC front-end component of speaker-recognition systems.

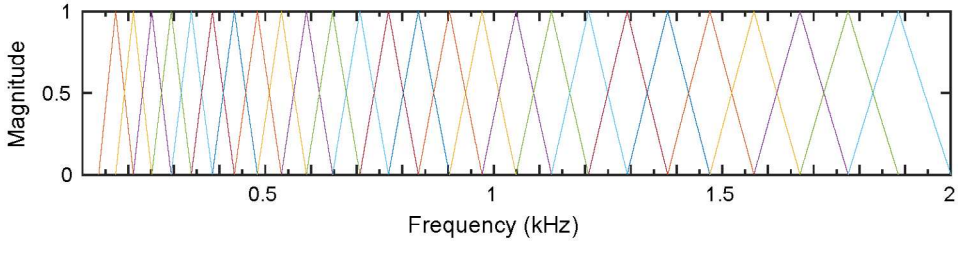


Figure 2. Example of a 27-channel mel-filterbank spanning the frequency range 100–2000 Hz.

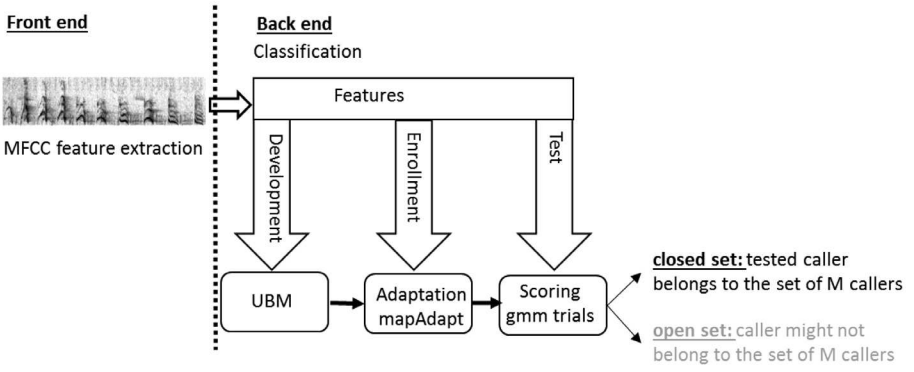


Figure 3. Two stages of speaker recognition system: MFCC extraction in the front end and classification procedure in the back end.

Belin (2006) argues that many of the "voice perception" abilities might be shared between human and non-human primates. Therefore, MFCC feature representations might be good candidates for caller recognition studies in non-human primates. Other reasons to use MFCCs in our Gaussian mixture model (GMM)-based framework exist as well. First, thanks to the discrete cosine transform used in their calculation, MFCCs tend to be uncorrelated and can thus be more efficiently modelled using GMMs with diagonal covariance matrices, as applied to text-independent individual recognition, which works regardless of the actual words being spoken. Second, noise robustness of the MFCC features has been demonstrated in several previous studies (Quatieri 2002; Clemins et al. 2005; Fox et al. 2008). Third, MFCCs show high accuracy, stability, and repeatability (Cheng et al. 2010). And finally, MFCCs are now also increasingly applied to caller recognition of animals such as elephants (Clemins et al. 2005), passerine birds (Trawicki et al. 2005; Fox 2008; Cheng et al. 2010), toothed whales (Brown et al. 2010), and blue monkeys (Mielke and Zuberbühler 2013).

As shown in Figure 3, speaker recognition systems consist of two stages, namely feature extraction (i.e. MFCCs as mentioned above), also called front-end, and classification, also called back-end, where acoustic space (i.e. distribution of acoustic features) for each speaker is estimated using GMMs (enrolment) and subsequent verification trials are used as validation (Sadjadi et al. 2013). Sadjadi et al. adopted a GMM-UBM framework, where the universal background model (UBM) is a GMM that is trained on a pool of data from a number of speakers (or here: callers). Ideally, the UBM should reflect actual operating acoustic conditions at hand (i.e. background noise). There are also speaker-specific (or caller-specific) models that are adapted from the UBM using the maximum a posteriori (MAP) estimation approach. After the enrolment phase, each test call is scored either against all enrolled speaker models to detect who is calling (speaker identification), or against the background model and a given speaker model to accept/reject an identity claim (speaker verification) (Reynolds et al. 2000). Here, two scenarios are possible: in the first case, this identity claim refers to a closed-set where the speaker or caller is known a priori to belong to a set of M speakers (callers). In the second case, a speaker or caller does not belong to the set, which is called an open-set scenario. In the latter case a threshold value is needed to decide whether a speaker (caller) belongs to the set or is out of the set (Ramachandran et al. 2002).

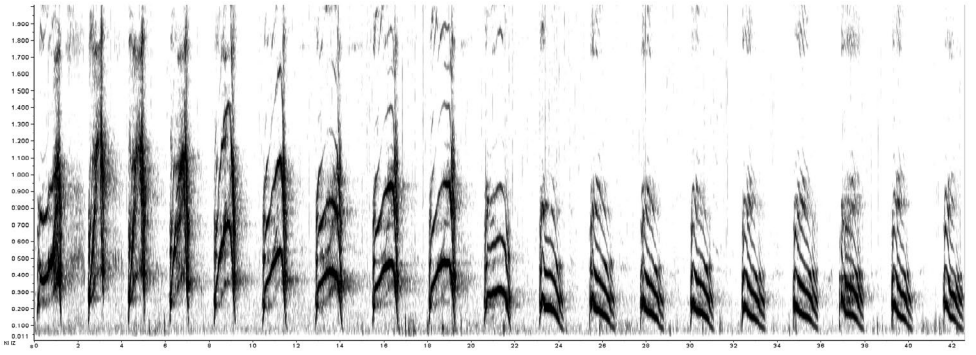


Figure 4. Spectrogram of a long call as used for MFCC extraction.

Almost all the previous studies on individual identification used MFCCs that were extracted from high-quality recordings with close-range microphone-caller distances. But in the case of an ALS with stationary microphones, microphone-caller distances, and thus the quality of the signals, are highly variable. In this study, we test the accuracy of an automatic caller recognition approach in orangutan (*Pongo pygmaeus wurmbii*) long calls (Figure 4). The long call is a series of pulses of different pulse types and of highly variable total duration, and discrimination has been reported on a geographic, individual, and contextual level (Delgado 2007; Ross and Geissmann 2007; Spillmann et al. 2010). We test two sets of long call recordings in order to match recording conditions (Ramachandran et al. 2002): (i) recordings recorded with short microphone-caller distance (10–20 m) using a Sennheiser ME67 directional microphone during individual focal follows to evaluate the feasibility of this approach in general, subsequently referred to as high-quality recordings, and (ii) recordings from a microphone grid (ALS), with lower signal-to-noise ratios (SNR) due to increased and variable microphone-caller distances and the omnidirectional characteristics of the SMX-II microphone, subsequently referred to as low-quality recordings. We test the validity of this procedure in such a system (more details on ALS see Spillmann et al. 2015).

A speaker recognition system can be text-dependent or text-independent. The difference between the two is that in the text-dependent system a fixed phrase needs to be spoken by each individual whereas in the text-independent system no restriction exists in this respect (Ramachandran et al. 2002). The GMM-UBM framework we employed is text-independent. We used this text-independent approach for two reasons: (i) the long calls produced by flanged male orangutans consist of various distinct pulse types (Ross and Geissmann 2007; Spillmann et al. 2010), and (ii) long call duration is highly variable ranging from 18 s up to 3 min (Delgado 2007). Previous work used acoustic measurements taken at particular pulse types (Delgado 2007; Spillmann et al. 2010) to demonstrate the feasibility of individual discrimination (not to be mistaken with automatic identification).

We approach the challenge of individual identification of long-calling orangutan males with a software originally developed for human speaker recognition. We used MFCCs for feature extraction available in MATLAB Voicebox (http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html) as front-end, and the GMM-UBM framework available in the MSR identity toolbox in MATLAB (Sadjadi et al. 2013) as back-end for speaker recognition in order to validate the potential of caller recognition in orangutans applied to a closed set of individuals.

Methods

Study area

Long call recording data presented here stem from Tuanan field station in Central Kalimantan, Indonesia (2.151° South; 114.374° East). The field site is part of an area managed by a non-governmental organization (Mawas) and covers a 1000 ha trail-system in a peat swamp forest that previously underwent selective logging 20 years earlier. Researchers and well-trained long-term field assistants conducted individual focal follows according to the standardized field methods (available online at http://www.aim.uzh.ch/Research/orangutannetwork/FieldGuidelines.html). This research project adhered to the American Society of Primatologists (ASP) principles for the ethical treatment of non-human primates.

Caller recognition

High-quality recordings

During individual focal follows of flanged male orangutans long calls were recorded whenever possible. We used a Sennheiser ME67 shotgun microphone in combination with a Roland Edirol R09 digital recorder. Recordings were taken at distances of around 10–20 m from the calling male with a sample rate of 44.1 kHz and sample size 16-bit. We tested 224 long calls given by 14 different individuals. Recordings were taken during 2007–2013. We tested the accuracy of caller recognition by using the MATLAB MSR Identity toolbox (Sadjadi et al. 2013). Because the frequency content of the recordings is limited to at most 3 kHz, the original recordings were re-sampled to 16 kHz to increase the computational efficiency of the subsequent processing for the caller identification task.

Three main steps are involved in the speaker recognition procedure: (1) extraction of MFCCs from all waveforms, (2) training of the recognition algorithm with GMMs (GMM-UBM framework), and (3) scoring each test call against all speaker models established in the training phase to determine the caller's identity. We divided the long call data into 4 different training sets (80% of an individual's calls) and 4 different test sets (20% of an individual's calls). Each established verification is scored and assigned to a probability value of matching the identified individual through the GMM-UBM framework.

We ran two different caller recognition tests with the sampled high-quality recordings. The first one included MFCC extractions established with mel frequency filters adjusted to the range 100–2000 Hz. The second test included MFCC extractions with mel frequency filters adjusted to the range 130–1400 Hz, reflecting the available frequency range of far-distance recordings where sound degradation acted on higher frequencies (see below ALS recordings). We aimed to test whether lower frequency ranges downgrade accuracy of caller recognition and if so to what extent.

To test whether the probability of identification of the GMM-UBM framework predicted identification, we ran a binomial generalized linear mixed model (GLMM) with correctness of recognition as response variable and the GMM-UBM probability of recognition as predictor, including individuals as random effect.

Recordings from the ALS

The microphone grid consisted of 20 omnidirectional SMX-II microphones in combination with SM2+ recorders and covered an area of 450 ha localization area at Tuanan field site

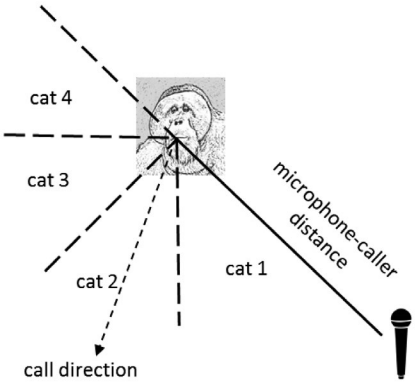


Figure 5. Microphone-caller distance and angle between call direction and microphone.
Notes: We used 4 angle categories. The example in the figure shows a call direction of angle category 2 (>45° <90°) in relation to the microphone position.

(Spillmann et al. 2015). Sample rate was set to 22.05 kHz but was subsequently downsampled to 16 kHz in order to increase the computational efficiency of the speaker recognition system. We tested long calls recorded by the ALS with known identity due to individual focal follow data where time and location of calling is recorded. We were able to establish an ALS data-set with long calls given by 10 individuals, with a total number of 123 long calls to test caller recognition with known microphone-caller distances. We selected long call recordings that showed the whole frequency range of the fundamental frequency in the spectrographic representations and without interfering background noise by conspecific (long call overlap) or heterospecific sounds (e.g. gibbon songs, pigtail macaque calls) but also noise of anthropogenic origin, such as the sound of boat engines. Microphone-caller distances varied between 20 and 420 m. As with the recordings made during focal follows, mel frequency filters were adjusted to the range 130–1400 Hz. Similar to our previous experiment, we divided the long call data into 4 different training sets (80% of an individual's calls) and 4 test sets (20% of an individual's calls). In order to visualize the performance of the caller recognition procedure we established a confusion matrix which contains information about the actual and predicted classification done by the GMM-UBM framework.

As with the focal animal recordings, we also used a binomial GLMM to test whether the probability of identification of the GMM-UBM framework, the microphone-caller distance and the angle between call direction and microphone position (see Figure 5) were valuable predictors for correct or incorrect identification. We used 4 categories for the angle-variable: 1 = long call direction <45°; 2 = LC direction >45° <90°; 3 = LC direction >90° <135°; and 4 = LC direction >135° <180°. We controlled for the possible impact of individuals by entering them as random effect.

Results

High-quality recordings

We used automatic speaker recognition software to establish the accuracy of orangutan long call identification of high-quality recordings with microphone-caller distances between 10 and 20 m. We ran the identification procedure two times with different filterbank

Table 1. Results of caller recognition using high-quality recordings from 14 individuals with 224 identifications.

	1. set	2. set	3. set	4. set	Average	Range
Frequency range (Hz)	% of correct assignments					
a. 100–2000	91.07	91.07	94.64	96.49	93.32	91.07–96.64
b. 130–1400	87.5	85.71	96.43	92.98	90.66	85.71–96.43

Table 2(a) and (b). The effect of the probability value (a) full bandwidth of long calls and (b) reduced bandwidth on correctness of caller identification.

	Estimate	Std. Error	*z* value	Pr(>|z|)
Dataset (a)				
(Intercept)	−17.145	3.863	−4.438	
Probability	33.68	7.026	4.794	<0.001
$\chi^2_{LRT} = 47.005, p < 0.001$				
Dataset (b)				
(Intercept)	−9.044	3	−3.015	
Probability	19.491	5.24	3.72	<0.001
$\chi^2_{LRT} = 19.838, p < 0.001$				

Note: Parameter estimates, associated standard errors and statistical significance as obtained from a binomial GLMM.

configurations for MFCCs. The frequency range 100–2000 Hz resembles the frequency distribution of a long call and with this range we obtained a 93.32% correct caller recognition rate. In order to test the accuracy of the smaller bandwidth that resembles the bandwidth of lower quality recordings, we extracted MFCCs between 130 and 1400 Hz of the same data-set as above and observed a slight degradation in accuracy with a 90.66% correct caller recognition rate (Table 1). Thus, there was a marginal reduction in accuracy of around 3% with the reduced bandwidth.

We examined with a binomial GLMM whether the probability value of the GMM-UBM framework is a valuable predictor for correctness of caller recognition. We corrected for individual impact by entering them as a random effect. Both statistical models with data-sets a (full bandwidth) and b (reduced bandwidth) showed that the probability value of the GMM-UBM framework is a highly significant predictor for correctness of caller identification (see Table 2(a) and (b)).

Recordings from the ALS

In the next step, we tested the accuracy of long call identification with variable microphone-caller distances (20–420 m). We ran the identification algorithm for 123 long calls given by 10 individuals, and found that 72.23% of calls were correctly identified (Table 3). The confusion matrix (see Table 4) of caller recognition shows the results for each tested individual for distances <420 m and distances <300 m, with the improvement of identification with shorter microphone caller distances indicated in bold. The result of the binomial GLMM shows that the probability value of the GMM-UBM framework best predicts correct or incorrect assignments (see Table 5) but also that increasing microphone-caller distance negatively affects identification whereas an increasing angle between call direction and microphone shows a tendency to do so too (see Figure 6 and Table 5).

Table 3. Results of caller recognition using long call recordings from the microphone grid with variable caller-microphone distances (20–420 m).

	1. set	2. set	3. set	4. set	Average	Range
Frequency range (Hz)	% of correct assignments					
130–1400	77.41	80.64	63.33	67.74	72.28	63.33–80.64

Note: Sample consisted of 123 long calls given by 10 individuals.

Table 4. Confusion matrix of caller recognition with variable microphone-caller distances.

		Classification										
		Chili	Dayak	Helium	Henk	Katman	Otto	Preman	Teju	Tomi	Wodan	% correct
Label	Chili	12/11	0	0	1/1	0	5/4	1/1	1	0	0	60.00/**61.11**
	Dayak	1	4/2	0	0	0	0	0	0	0	0	80.00/66.67
	Helium	1	0	13/12	0	0	3/2	0	0	0	0	76.47/**80.00**
	Henk	1	0	1/0	3	0	1/0	0	1/0	0	0	42.86/**75.00**
	Katman	0	0	0	0	7	1	0	0	1	0	77.78
	Otto	1	0	0	0	0	8/7	0	0	1	0	80.00/77.78
	Preman	0	0	0	0	0	2/1	5	1	0	1/0	55.56/**80.95**
	Teju	1	0	0	0	0	2	1	17	0	0	80.95
	Tomi	0	0	0	0	1	0	0	0	8	2/1	72.73/**80.00**
	Wodan	0	0	0	0	0	0	1	1	0	12	85.71

Note: Improvement of caller recognition by reducing microphone-caller distance from <420 to <300 m denote in bold.

Table 5. The effect of the probability value, the microphone-caller distance, and the angle calling direction-microphone on correctness of caller identification.

Fixed effects	Estimate	Std. Error	*z* value	Pr(>|z|)
(Intercept)	−24.317	7.077	−3.436	
Probability	46.317	13.126	3.529	<0.001
Microphone-caller distance	−0.760	0.321	−2.369	0.018
Angle calling direction-microphone	−1.279	0.674	−1.897	0.058
$\chi^2_{LRT} = 32.63$, $p < 0.001$				

Note: Parameter estimates, associated standard errors and statistical significance as obtained from a binomial GLMM.

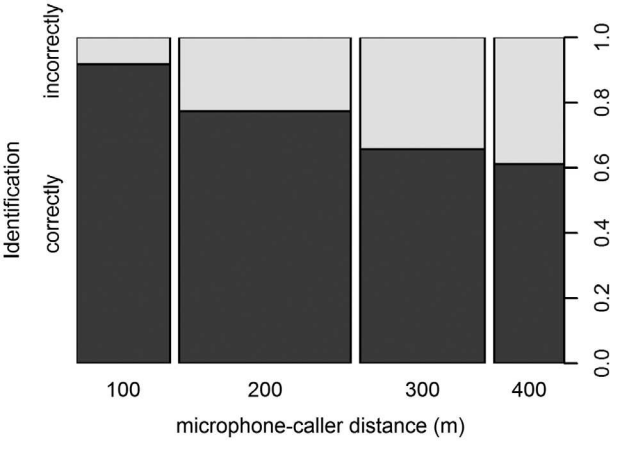


Figure 6. Likelihood of correct identification with increasing microphone – caller distances.

Discussion

The aim of this study was to test the potential of an automatic speaker recognition system (originally developed for human voice recognition) applied to orangutan long calls. High-quality recordings with close-range microphone-caller distance (directional microphone) yielded 93.3% (by chance 7.14%) correct caller recognition (Table 1). Low-quality recordings (omnidirectional microphone) and variable, usually far greater, microphone-caller distances (20–420 m) yielded 72.28% (by chance 10%) correct caller recognition (Table 3). These results reflect the ability of individual recognition known from experimental tests in different primate species (i.e. Snowdon and Cleveland 1980; Cheney and Seyfarth 1982; Rendall et al. 1996). In both cases (high- and low-quality recordings), the probability value of the GMM-UBM framework can be used as a benchmark for correctness of caller recognition. This result indicates that high probability values indicate more confidence in accuracy of the identification result (Tables 2 and 5). As expected, microphone-caller distance had a negative effect on the accuracy of caller recognition, whereas the angle between call direction and microphone tended to do this too (Table 5 and Figure 6).

The degradation of accuracy in caller recognition from high-quality recordings to low-quality recordings was expected, but despite the large variation of microphone-caller distances the result of the low-quality recordings (72.28%) is still far above the identification success expected by chance of 10% (for 10 individuals). The different microphone-caller distances incorporate much variation based on biotic and abiotic factors. Biotic factors include background noise caused by other animal calls or by anthropogenic origin (noise of boat engines), and degradation of sounds by the vegetative structure of the location. Abiotic factors include attenuation of sounds due to the distance of the recorded call, or weather- or climate-related effects (Kirschel et al. 2009). A major limitation in applying the validated method is distance. Our result suggests that long call identification at far distances yields lower accuracy due to sound attenuation and signal degradation that results in low SNR. Nevertheless, our approach yields a remarkably high accuracy in light of the highly variable microphone-caller distances. Indeed, another study that tested individual discrimination of long calls by applying acoustic measurements of re-recorded playbacks at different distances, found possible discrimination up to 300 m (Lameira and Wich 2008). Additionally we found a tendency of a negative effect of the angle between call direction and microphone, which implies that long calls may have a directional characteristic.

Our results suggest that automatic individual identification is reliable for distances up to 420 m. The limitations of this approach are sound degradation and background noise that masks the signal of interest. This raises the question whether these processes also impair the orangutans' ability to recognize the identity of the long caller at farther distances. Evidence for improved perception of a target signal in spite of noise and degradation comes from the so-called cocktail party effect encountered in human and animal communication literature, where selective attention tunes out all but the signal of interest (Aubin and Jouventin 1998; Ebata 2003). Obviously, group living animals may face the cocktail party problem, but also heterospecific signallers may contribute to a receiver's perception problem (Bee and Micheyl 2008). In orangutans the latter incidence is plausible: orangutans are semi-solitary and long call overlap is rare, but rainforests are quite noisy habitats (Slabbekoorn 2004). In future work, we plan to test whether responses to distant long calls deviate from those to closer ones.

In conclusion, this study demonstrates that an automatic speaker recognition system originally developed for human voice recognition provides a promising and viable tool for animal vocal communication research and in particular for the identification of calling individuals recorded by a PAM system. Note, however, that the validation presented here considered a closed-set scenario, and that when applying this method to PAM data without knowledge of the caller identity, an open-set approach is required to account for new individuals that may move into a monitored area. We propose that automatic caller recognition is a valuable tool to identify individuals by their calls stemming from PAM and ALSs. Its use can greatly facilitate field research on vocal communication.

Acknowledgements

We gratefully thank the Indonesian Institute of Science (LIPI), the Indonesian State Ministry for Research and Technology (RisTek), the Kementerian Kehutanan (Direktorat Jenderal PHKA), Kementerian Dalam Negeri, the local government in Central Kalimantan, the BKSDA Palangkaraya, the Bornean Orangutan Survival Foundation (BOSF), and MAWAS in Palangkaraya for their permission and support to conduct this study. We also thank Universitas Nasional (UNAS) for support and collaboration and particularly Sri Suci Utami. Thanks go to our local field team in Tuanan; Pak Rahmadt, Tono, Idun, Abuk, Kumpo, Suwi, Suga, and Ibu Fitriah. Additionally we thank all local and foreign students and researchers for their contribution in data collection. We gratefully thank Maria van Noordwijk for directing the Tuanan field project. Many thanks go to Mure Wipfli, for technical support and maintenance of the passive acoustic monitoring and Martina Nueesch for her effort in data extraction. We also thank the anonymous reviewers for comments and suggestions on this manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Leakey Foundation [grant number 353]; Basler Stiftung für biologische Forschung; Foundation/Stiftung Dr. Joachim de Giacomi; A. H. Schultz Stiftung/Foundation.

ORCID

Brigitte Spillmann http://orcid.org/0000-0002-0694-3670
Carel P. van Schaik http://orcid.org/0000-0001-5738-4509
Tatang M. Setia http://orcid.org/0000-0003-2667-5199
Seyed Omid Sadjadi http://orcid.org/0000-0002-1654-733X

References

Aubin T, Jouventin P. 1998. Cocktail-party effect in king penguin colonies. Proc R Soc Lond [Biol]. 265(1406):1665–1673.

Bee MA, Kozich CE, Blackwell kJ, Gerhardt HC. 2001. Individual variation in advertisement calls of territorial male green frogs, rana clamitans: implications for individual discrimination. Ethology. 107(1):65–84.

Bee MA, Micheyl C. 2008. The "cocktail party problem": what is it? How can it be solved? And why should animal behaviorists study it? J Comp Psychol, 122(3), 235–251.

Belin P. 2006. Voice processing in human and non-human primates. Philos Trans R Soc Lond B Biol Sci. 361(1476):2091–2107.
Brown JC, Smaragdis P, Nousek-McGregor A. 2010. Automatic identification of individual killer whales. J Acoust Soc Am. 128(3):EL93–EL98.
Cheney DL, Seyfarth RM. 1982. Recognition of individuals within and between groups of free-ranging vervet monkeys. Am Zool. 22(3):519–529.
Cheng J, Sun Y, Ji L. 2010. A call-independent and automatic acoustic system for the individual recognition of animals: a novel model using four passerines. Pattern Recogn. 43(11):3846–3852.
Clemins PJ, Johnson MT, Leong KM, Savage A. 2005. Automatic classification and speaker identification of African elephant (Loxodonta africana) vocalizations. J Acoust Soc Am. 117(2):956–963.
Davis S, Mermelstein P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process. 28(4):357–366.
Delgado RA. 2007. Geographic variation in the long calls of male orangutans (*Pongo* spp.). Ethology. 113(5):487–498.
Ebata M. 2003. Spatial unmasking and attention related to the cocktail party problem. Acoust Sci Technol. 24(5):208–219.
Ehnes M, Foote JR. 2015. Comparison of autonomous and manual recording methods for discrimination of individually distinctive Ovenbird songs. Bioacoustics. 24(2):111–121.
Fitch T. 2006. Production of vocalizations in mammals A2. In: K Brown. Encyclopedia of language & linguistics. 2nd ed. Oxford: Elsevier; p. 115–121.
Fox EJS. 2008. A new perspective on acoustic individual recognition in animals with limited call sharing or changing repertoires. Anim Behav. 75(3):1187–1194.
Fox EJS, Roberts JD, Bennamoun M. 2008. Call-independent individual identification in birds. Bioacoustics. 18(1):51–67.
Galeotti P, Sacchi R. 2001. Turnover of territorial Scops Owls Otus scops as estimated by spectrographic analyses of male hoots. J Avian Biol. 32(3):256–262.
Gilbert G, Tyler GA, Smith KW. 2002. Local annual survival of booming male Great Bittern Botaurus stellaris in Britain, in the period 1990–1999. Ibis. 144(1):51–61.
Kirschel ANG, Earl DA, Yao Y, Escobar IA, Vilches E, Vallejo EE, Taylor CE. 2009. Using songs to identify individual mexican antthrush formicarius moniliger: comparison of four classification methods. Bioacoustics. 19(1–2):1–20.
Lameira AR, Wich SA. 2008. Orangutan long call degradation and individuality over distance: a playback approach. Int J Primatol. 29(3):615–625.
Mielke A, Zuberbühler K. 2013. A method for automated individual, species and call type recognition in free-ranging animals. Anim Behav. 86(2):475–482.
Peake TM, McGregor PK. 2001. Corncrake Crex crex census estimates: a conservation application of vocal individuality. Anim Biodivers Conser. 24(1):81–90.
Quatieri TF. 2002. Discrete-time speech signal processing: principles and practice. Upper Saddle River (NJ): Prentice Hall.
Ramachandran RP, Farrell KR, Ramachandran R, Mammone RJ. 2002. Speaker recognition—general classifier approaches and data fusion methods. Pattern Recogn. 35(12):2801–2821.
Rendall D, Rodman PS, Emond RE. 1996. Vocal recognition of individuals and kin in free-ranging rhesus monkeys. Anim Behav. 51(5):1007–1015.
Reynolds DA, Quatieri TF, Dunn RB. 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing. 10(1–3):19–41.
Ross MD, Geissmann T. 2007. Call diversity of wild male orangutans: a phylogenetic approach. Am J Primatol. 69(3):305–324.
Sadjadi SO, Slaney M, Heck L. 2013. MSR Identity Toolbox v10: A MATLAB Toolbox for Speaker-Recognition Research. Speech and Language Processing Technical Committee Newsletter. 1(4).
Slabbekoorn H. 2004. Habitat-dependent ambient noise: Consistent spectral profiles in two African forest types. J Acoust Soc Am. 116(6):3727–3733.
Snowdon CT, Cleveland J. 1980. Individual recognition of contact calls by pygmy marmosets. Anim Behav. 28(3):717–727.

Spillmann B, Dunkel LP, van Noordwijk MA, Amda RNA, Lameira AR, Wich SA, van Schaik CP. 2010. Acoustic properties of long calls given by flanged male orang-utans (*Pongo pygmaeus wurmbii*) reflect both individual identity and context. Ethology. 116(5):385–395.

Spillmann B, van Noordwijk MA, Willems EP, Mitra Setia T, Wipfli U, van Schaik CP. 2015. Validation of an acoustic location system to monitor Bornean orangutan (*Pongo pygmaeus wurmbii*) long calls. Am J Primatol. 77(7):767–776.

Taylor AM, Reby D. 2010. The contribution of source–filter theory to mammal vocal communication research. J Zool. 280(3):221–236.

Tibbetts EA, Dale J. 2007. Individual recognition: it is good to be different. Trends Ecol Evol. 22(10):529–537.

Townsend SW, Charlton BD, Manser MB. 2014. Acoustic cues to identity and predator context in meerkat barks. Anim Behav. 94:143–149.

Trawicki MB, Johnson MT, Osiejuk TS (2005). Automatic Song-Type Classification and Speaker Identification of Norwegian Ortolan Bunting (Emberiza Hortulana) Vocalizations. 2005 IEEE Workshop on Machine Learning for Signal Processing, Mystic, Connecticut; pp. 277–282.

Wich SA, Koski S, de Vries H, van Schaik CP. 2003. Individual and contextual variation in Thomas langur male loud calls. Ethology. 109(1):1–13.

Xia C, Lin X, Liu W, Lloyd H, Zhang Y. 2012. Acoustic identification of individuals within large avian populations: a case study of the brownish-flanked bush warbler, South-Central China. PLoS One. 7(8):e42528.